ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Special Issue, February 2017, Volume 5, Issue 2,

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Survey of Machine Translation Techniques

Shantanoo Dubey
Department of Computer Science and Engineering
Jabalpur Engineering College,
Madhya Pradesh – India

Abstract: Natural Languages gives us the flexibility to express a given concept or thing in a variety of ways. Machine translation is the process of automatically translating text or speech from one natural language into another. With rule-based, data-driven and hybrid-based techniques machine translation can be used to translate confidential documents, removal of language barrier in terms of education and learning, etc. Although various techniques for machine translation are proposed, but none of them is the state-of-the-art, which can be used effectively for every language. Thus in this paper, we aim to present a systemic literature review of the existing techniques for machine translation, along with some important issues in this domain. So that this paper may help in future research and in choosing an adequate technique for a particular purpose.

Keywords: Machine translation, Rule-based machine translation, Data-based machine translation, Hybrid-based machine translation.

I. Introduction

Machine Translation (MT) is the computerized process of converting a piece of text from a source language(SL) to a target language(TL) while ensuring the original meaning of the source text[1] In this technically integrated world, a person can easily access a huge volume of multi-lingual content in the common platforms like internet, but the problem comes when the content is written in the language which is not recognizable by a user. So, to solve this problem there is a requirement of some translating medium, which may be human translators, but they are limited and expensive resource in this competitive world. Thus, fully automated translation is emerged as a viable tool to solve language translation problems and comes in the focus of linguists, computer scientists and engineers to propose improvements in the translation. Consequently fully automated translation comes with three major techniques viz.: rule-based MT, data-driven translation and hybrid translation technique.

For understanding, we divide evolution of MT techniques in three phases. The first phase came with rule-based MT technique which was used in systems like Systran [2] during 1970. In the second phase, an approach of data-driven MT i.e., statistical machine translation was introduced for commercial use by Thomas J. Watson Research Center [3]. In the last phase, hybrid machine MT was proposed, with the aim of taking advantages of former two techniques. However, despite of the efforts, translation quality by any single technique is still poor as compare to that of human translation. Thus, considering its vitality in digital world, a systematic review is required because each technique differs in terms of analytical power to analyze SL and the level up to which technique is able to preserve the meaning of SL in the TL. Furthermore, the correct choice of translation technique is also important because the cost of translation depends on the linguistic knowledge and theoretical framework required for the system, which varies with each technique.

In this paper, we present a systematic literature review of the techniques used for MT along with their pros and cons. We have also mentioned recent work that had been performed by various scholars to increase the efficiency of translation in proposed techniques. Following the introduction, this paper is organized as follows: Section 2 includes search strategy; Section 3 reveals related work in this domain; Section 4 briefs about the proposed techniques and lastly, Section 5 draws the conclusion.

II. Search Strategies

This section gives an overview of our strategy towards searching for the details on this topic. Electronically, we referred to various textbooks on machine translation, papers published in conferences and journals pertaining to fully automated translation. We read paper's title, keywords and it's abstract to find its relevance with our topic. Further, we read the contents in details for those which gave any clue or information about any machine translation technique. We performed electronic search within five electronic databases: IEEE Xplore, ACM Digital Library, Elsevier, Compendex and SpringerLink.

We also referred the textbooks like Statistical Machine Translation by Koehn Philip, Example-based machine translation by Sato, Satoshi. Moreover, we searched papers in various summits and conferences like IEEE international Conferences, Association for Computational Linguistic's annual meeting, International Conference on Computational Linguistics, International Conference on Natural Language Processing and Machine Translation Summits.

Manually we searched papers published in potentially relevant, peer- reviewed journals: IEEE Transaction on Audio, Speech and Language Processing, Computational Linguistics, International Journal of Translation, Natural Language Semantics, Transaction on Speech and Language Processing, Computing and Informatics, Natural Language Engineering, Computer Speech & Language, Computational Intelligence & Communication Technology, Language Resources and Evaluation, Machine Translation.

III. Related Work

The need of literature review in the MT techniques was already felt by research community in mid-eighties, and Slocum [4] published a survey paper titled "A survey of machine translation: its history, current status, and future prospects" in 1985. The author explained the research work in MT, and in parallel he also supported the idea of future development in multilingual analysis and synthesis tools. Then, Hutchins [5] came with the brief review of work in MT up to 1988. Subsequently, in the emerging light of SMT, Dorr, Jordan, and Benoit [6] surveyed the MT research work centered in US, Europe and Japan in his paper titled- A survey of current paradigms in machine translation. Later in 2010, Tripathi and Sarkhel [7] proposed the superficial view over techniques and revealed that lack in universality in lexicon, ambiguity and linguistic irregularities degrades the quality of translation. Another study was proposed by Antony[8] in his paper titled "Machine Translation Approaches and Survey for Indian Languages" at 2013, he put major emphasis on development of MT for indian languages. In 2014, Okpor[9] surveyed over the issues and challenges of MT techniques. Furthermore, Kituku, Muchemi, and Nganga [10], proposed the review of MT techniques, but they lack in covering proposed research work in realm of MT techniques.

IV. Machine Translation Techniques

In this section, we provide a superficial view of proposed MT techniques. MT can be achieved either by using two languages, which is called bilingual translation or by using more than two languages, which is called multilingual translation. While translating any sentence, system adopts meta-phrase and/or para-phrase for the translation. Meta phrase involves word by word translation from SL to TL without focusing on the meaning of the original text. Whereas in para phrase, the syntactic order of words may not be preserved, but translated text must reflect the meaning of original text[7]. Thus, most of the current system refers para phrased translation rather than meta phrase translation. Generally, machine translation systems are classified into Human Translation with Machine Support, Machine Translation with Human Support and Fully Automated Translation. This paper focuses only on Fully Automated Translation which covers: rule-based translation, data-driven translation and hybrid-based translation. At present, most of the MT related research is based on data-driven and hybrid based techniques.

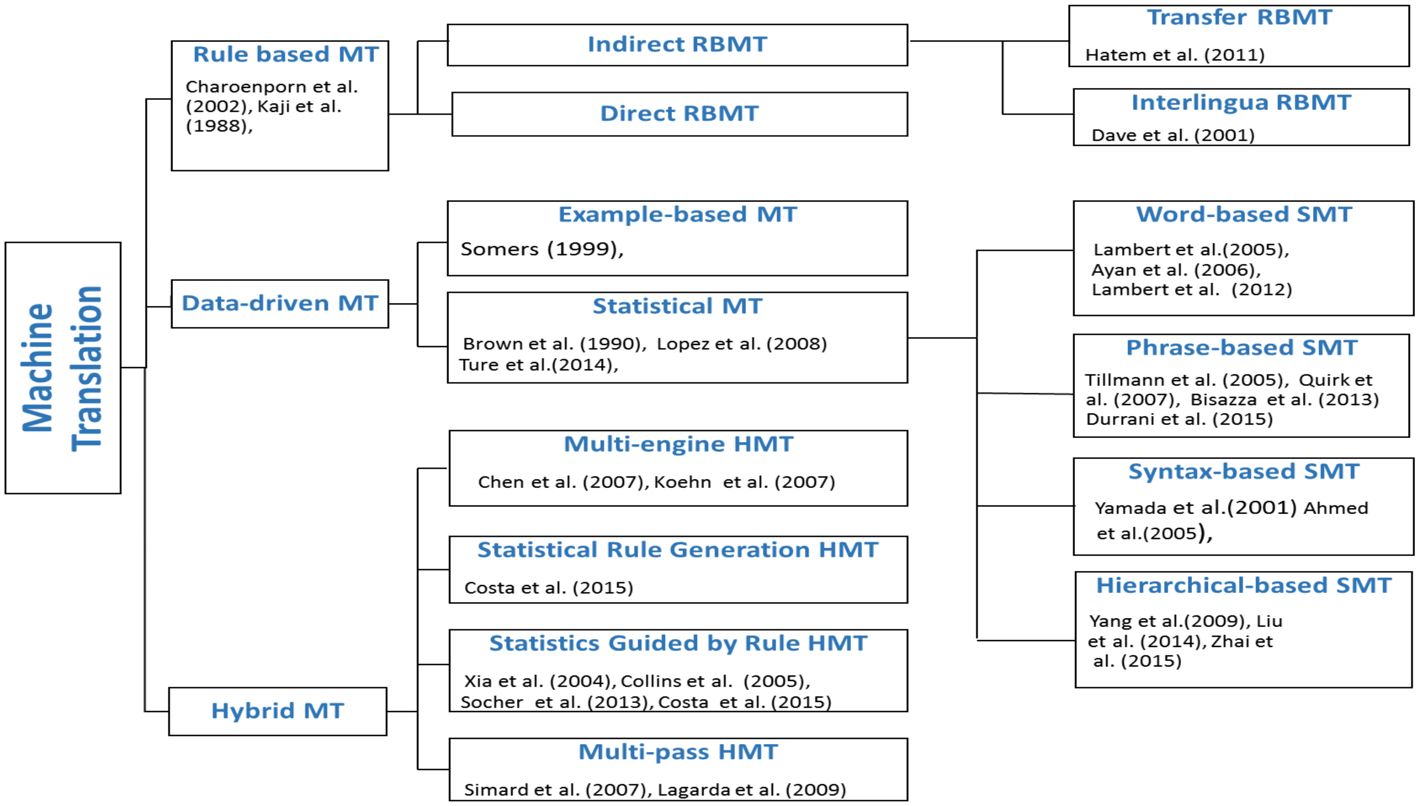
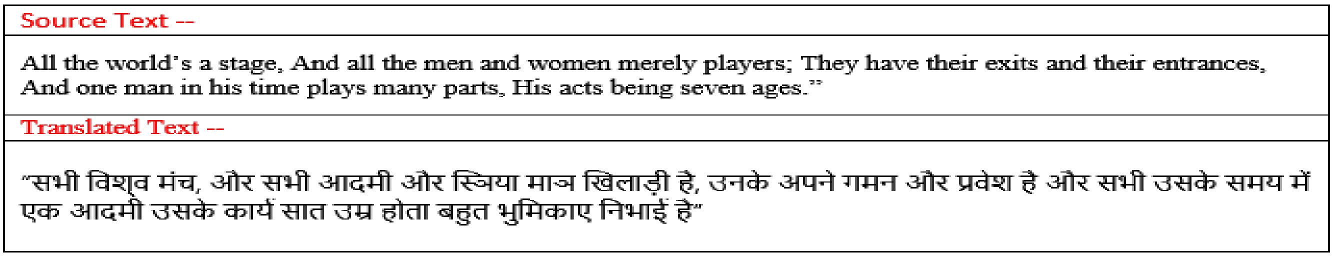

Machine Translation
Rule based MT
Charoenporn et al. (2002), Kaji et al. (1988),
Indirect RBMT
Direct RBMT
Transfer RBMT
Hatem et al. (2011)
Interlingua RBMT
Dave et al. (2001)
Data-driven MT
Example-based MT
Somers (1999),
Statistical MT
Brown et al. (1990), Lopez et al. (2008) Ture et al.(2014),
Word-based SMT
Lambert et al.(2005), Ayan et al. (2006), Lambert et al. (2012)
Phrase-based SMT
Tillmann et al. (2005), Quirk et al. (2007), Bisazza et al. (2013) Durrani et al. (2015)
Syntax-based SMT
Yamada et al.(2001) Ahmed et al.(2005),
Hierarchical-based SMT
Yang et al.(2009), Liu et al. (2014), Zhai et al. (2015)
Hybrid MT
Multi-engine HMT
Chen et al. (2007), Koehn et al. (2007)
Statistical Rule Generation HMT
Costa et al. (2015)
Statistics Guided by Rule HMT
Xia et al. (2004), Collins et al. (2005), Socher et al. (2013), Costa et al. (2015)
Multi-pass HMT
Simard et al. (2007), Lagarda et al. (2009)

Fig. 1: Classification of MT Techniques

A. *Rule Based Machine Translation (RBMT)*

The rule-based technique is the earliest proposed approach in the field of MT. In RBMT, sentences are translated by grammar rules, which are written on the ground of linguistic knowledge gathered from linguists. Thus, morphological, syntactic, and semantic analysis of both the SL and the TL are the cornerstones for the translation in RBMT. In the translation process of RBMT, linguistic rules are applied in three different phases : analysis phase, transfer phase and generation phase, which necessitate the use of syntax analysis, semantic analysis, syntax generation and semantic generation for translation[9]. Below fig. 2 is the snapshot of Anusaarka (http://anusaaraka.iiit.ac.in) for the RBMT translation of an English sentence taken from William Shakespeare's poem \The Seven Ages of Man", into a corresponding Hindi sentence.



Source Text --

All the world's a stage, And all the men and women merely players; They have their exits and their entrances, And one man in his time plays many parts, His acts being seven ages."

Translated Text --

"सभी विश्व मंच, और सभी आदमी और स्त्रिया मात्र खिलाड़ी है, उनके अपने गमन और प्रवेश है और सभी उसके समय में एक आदमी उसके कार्य सात उम्र होता बहुत भुमिकाए निभाई है"

Fig. 2: Snapshot of Anusaarka

For such translation, system/tool based on RBMT requires following minimum number of resources--

1. A dictionary that will map every word of SL to an appropriate word in TL.
2. Rules representing regular sentence structure of SL.
3. Rules representing regular sentence structure of TL.
4. Rules to establish synchronization between these two structures.

The RBMT's efficiency of execution can be increased by pre-analyzing the grammar to determine an antecedent set for each rule[12], this ensures that rule is active only when an action in the antecedent set for the rule is performed. Moreover, by using machine learning method- RIPPER[11] the quality of translation in RBMT will be improved significantly and RIPPER does not require linguistic knowledge for translation. There are two different approaches in the RBMT viz Direct RBMT and Indirect RBMT. Moreover, indirect RBMT is further divided into Transfer-Based approach and Interlingua Machine Translation Approaches(as shown in Vauquois Triangle(fig. 3)). Although all these belong to RBMT, but they vary in the depth of analysis of SL and the level to form language-independent representation of meaning between SL and TL.

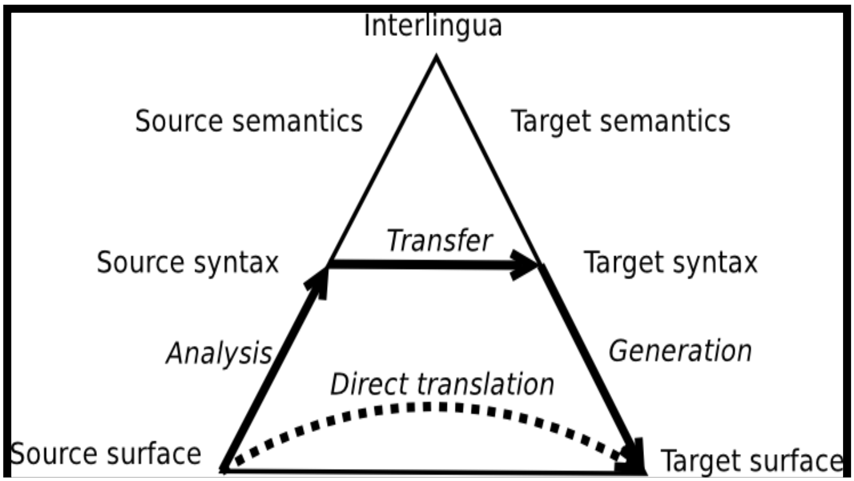


Fig. 3: Vauquois Triangle[https://noramachinetranslation.files.wordpress.com/2015/02/pyramid.png]

Direct Rule-Based Machine Translation: As direct RBMT or dictionary-based translation present at the lowest level of the vauquois triangle(fig. 3), words of the SL is translated directly into the TL without generating any intermediate presentation. It is the oldest and the less popular approach. This technique requires following four steps for generation of a sentence in TL[13].

1. Morphology analyzer identifies words of SL by removing ambiguities present in the sentence.
2. Bilingual dictionary generates TL base forms equivalent to SL base forms.
3. Then, TL word order is subjected to grammatical scrutiny with the help of rules defined on the basis of linguistic knowledge.
4. Lastly, the output is generated in translation language.

Thus it requires little syntactic and semantic analysis, and its performance depends on morphological analysis, text processing software, and word-by-word translation with minor grammatical adjustments on word order and morphology. However, linguistic and computational naivety is the main issue of this approach.

In-direct Rule-Based Machine Translation: In this multi-lingual translation technique, a less language-specific representation, termed as abstract, is generated on the basis of Morphology, semantic and syntactic analysis of SL. Then, an equivalent abstract is generated in TL with the help of specific generator (Fig. 3). Indirect RBMT is further divided into Transfer-Based and Interlingua Machine Translation Approaches.

Transfer-Based RBMT: It is a multi-lingual technique which is more common than other RBMT approaches. The distinguishing feature of transfer based RBMT is that it does not depends completely on the language pair involved in the translation. It breaks the process of translation in following three stages:--

1. Analysis stage, the syntactic representation is generated on the basis of morphology, syntax and semantics analysis of the SL with the help of SL lexicons and grammar.
2. Transfer stage, the SL intermediate structure is transferred to TL intermediate form, with the help of bilingual dictionary conversion.
3. Generation stage, the TL text is generated from the TL intermediate form with the help of TL grammar.

With such sequence of translation, it has some challenges like complexity of transfer module, work in reusable modules of analysis and synthesis. However, Hatem, Omar, and Shaker [14] proposed the morphological and syntactic symmetry for English and Arabic language for improving the translation quality of it. This technique has a potential to generate a result with 90% accuracy. Currently, Mantra, PaTrans & Vyakarta translation systems are on based transfer model.

Interlingua-based RBMT: Within the RBMT paradigm, the interlingual approach is present as an alternative to the direct approach and transfer approach. This approach transform the SL into an interlingual representation, which is a language independent representation. Then, target sentence is generated from language independent representation with the help of TL dictionaries(fig. 6). Thus, for translation this approach requires dictionaries (or lexicons) for analysis and generation (specific to the domain and the languages involved),a conceptual lexicon (specific to the domain), which is the knowledge base about events and entities known in the domain a set of projection rules (specific to the domain and the languages), grammars for the analysis and generation of the languages involved. The advantage of this approach is that language independent representation can be used to generate translation for different TLs. Moreover, this approach also provides an economical way to make multilingual translation systems because it creates only 2n pairs between the n language and interlingua, rather than n(n-1) language pairs in other approaches. However, the main disadvantage is that it fails in utilizing similarities between languages. Currently, UniTran translation model is using this approach, and it is preferable in circumstance where multiple languages are involved in translation.[15]

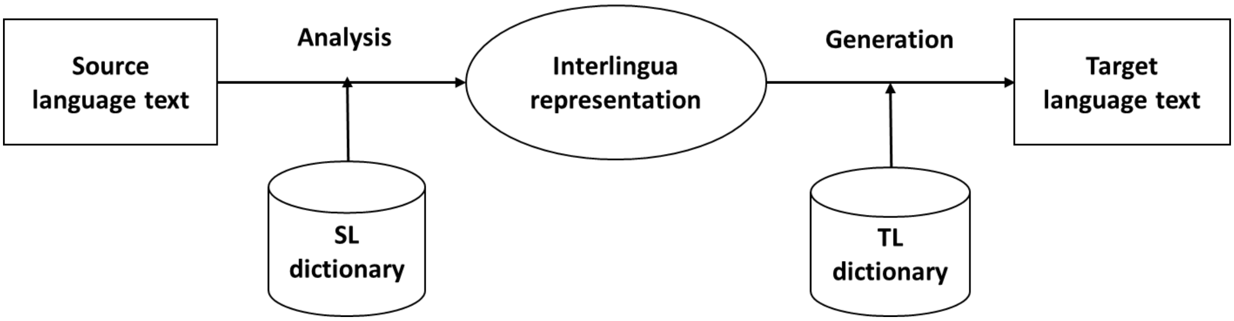


Fig. 6: Interlingua-based RBMT

Advantages of RBMT:-

1. Quality of RBMT is predictable and consistent.
2. RBMT is faster to update, maintain (can be done daily or more frequently).
3. RBMT requires knowledge and relatively expert human labor, but not much data.

Disadvantages of RBMT:-

1. Requirement of good bilingual dictionaries increases developmental and customization costs.
2. It does not handle ambiguous words and phrases easily.
3. It is hard to implement for big systems and expressions.

TABLE I Comparison of Direct RBMT, Transfer RBMT And Interlingua RBMT

	Direct RBMT	Transfer RBMT	Interlingua RBMT
Depth of analysis of SL	Without comprehensive analysis, target sentence is generated by direct word-by-word translation.	In this approach, an abstract is formed after morphological, syntax and semantics analysis of source text. But it lacks in lexical analysis.	Its analysis level is same as that of transfer RBMT
Language independent representation	language independent representation is not formed, during translation	Comparatively less language independent representation is formed, during translation	Completely language independent representation is formed, during translation.

B. Data-driven approach of machine translation (DDMT)

This method of translation has emerged as one of the widely explored areas of MT since 1990's, reason being the bilingual parallel aligned corpora which is used for the translation. Bilingual parallel aligned corpora is data having text and its translation, thus this technique is also called corpus-based machine translation. This technique first uses annotation process for the alignment of a parallel corpus and then creates classifier by either supervised, semi- supervised or unsupervised learning methods[16] . The main challenge for DDMT is the requirement of a parallel corpus for the translation, which may not be always present. Below is the comparison between RBMT and DDMT. The DDMT is classified into two techniques: statistical machine translation (SMT) and Example-based machine translation (EBMT).

Table II Comparison of RBMT and DDMT

S.No.	RBMT	DDMT
1	It is a rational approach.	It is an empirical approach.
2	It is generated on the basis of morphological, syntactic, and semantic analysis of SL and TL	It is generated by the analysis of a bilingual text
3	Adding new rules in RBMT system is quite difficult.	Adding new bilingual data is easy, and improves the ability of system for better translation.

Statistical machine translation (SMT): Initially, SMT was first introduced by Warren Weaver in 1949, but after some time, in the late 1980s, it was re-introduced by researchers at IBM's Thomas J. Watson research center[3]. SMT can be described as the process of finding and matching identical pairs from SL and TL in parallel corpora. The goal of SMT is to make an optimal decision in language translation by using statistical decision theory, based on probability distribution function. The important feature of SMT is the presence of statistical table, which can be built by using supervised or unsupervised statistical machine learning algorithms. Statistical table generally contains statistical information pertaining to sentences or languages. SMT relies on a statistical calculation of the probabilities of a match[17] by using two probabilistic models: Language model and Translation model, rather than relying on linguistic translation algorithms. The idea of SMT is that document can be translated on the basis of probability distribution function P(t/s), where P(t/s) is the probability of translating a sentence, say 's' in SL to a sentence 't' in TL. And this function is generated easily by using Bayes theorem. In Bayes theorem probability distribution p(t/s) is obtained from the product of P(s/t) and p(t), where P(s/t) is the probability that the source sentence is a translation of the target sentence, and P(t) is the probability of the TL.

In the architecture of SMT shown in fig. 7, three components play dominant role namely- language model, translation model and decoder[10]. First, the Language model calculates P(t) based on the monolingual corpus. It is also responsible for the correct combination of words in the TL. Thus, it ensures the grammatically correct output. Second, the Translation model, that calculates P(s/t) based on the parallel corpus by ensuring that machine translation system produce target hypothesis corresponding to the source sentence. Third, the Decoder, that performs the actual translation. Given a sentence t in the TL, the decoder chooses a viable translation by selecting sentence of maximum probability in the SL on the basis of the equation given below:-

$$P\ (t{=}s) = argmax(P\ (s{=}t)) * P\ (t))$$

With this architecture, following issues must be taken care of while designing an SMT model[9]--

1. Alignment: In parallel corpora, it may be possible that single sentence in one language has multiple translated sentences in another language. Therefore, Gale-Church alignment algorithm is preferred for better alignment of sentences.

2. Statistical Anomalies: Perhaps real-world training sets outweighs the quality of translation. Like a sentence "I ordered pasta in Domino's" is mistranslated into "I ordered pizza in Domino's" because of copiousness of "pizza in Domino's" in the training set.

3. Idioms: Depending on the corpora used, idioms may not translate "idiomatically".

4. Out of vocabulary (OOV) words: These words come into picture due to lack of parallel corpora. This scarcity varies in different domain depending on the SL and TL. This problem widens if domain is shifted from one to another. Paraphrasing was proposed as a solution to OOV words problem in MT.

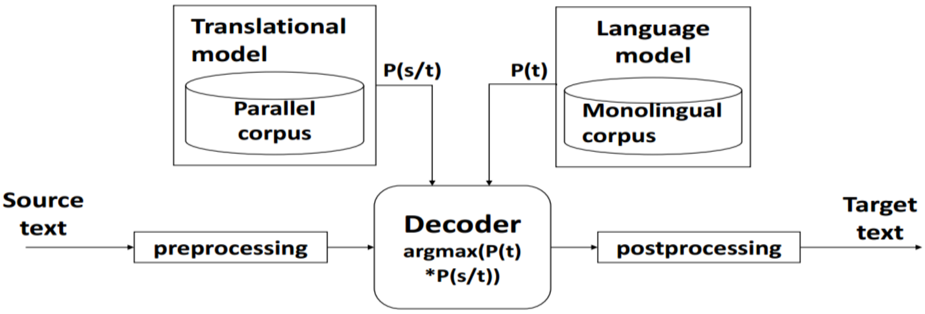
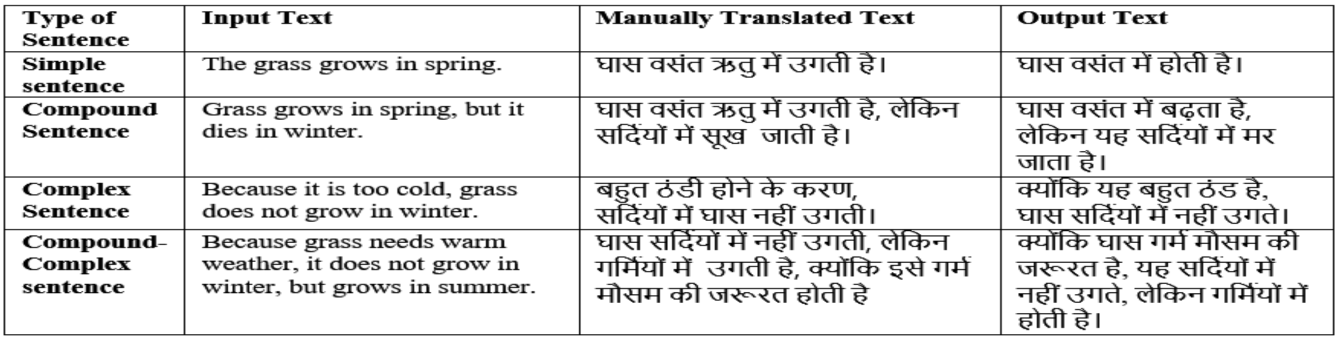


Fig. 7: Statistical Machine Translation

In fig. 8, we provide translations of Google translator(https://translate.google.co.in/), which uses SMT for translation, in the figure we mentioned the translation of different type of English(SL) sentences into a corresponding Hindi(TL) sentences. We choose these four sentences because they are explained as four major type of the English sentences in online English lectures of Stanford University. Moreover, figure also compares machine translated text from manually translated text.



Type of Sentence	Input Text	Manually Translated Text	Output Text
Simple sentence	The grass grows in spring.	घास वसंत ऋतु में उगती है।	घास वसंत में होती है।
Compound Sentence	Grass grows in spring, but it dies in winter.	घास वसंत ऋतु में उगती है, लेकिन सर्दियों में सूख जाती है।	घास वसंत में बढ़ता है, लेकिन यह सर्दियों में मर जाता है।
Complex Sentence	Because it is too cold, grass does not grow in winter.	बहुत ठंडी होने के करण, सर्दियों में घास नहीं उगती।	क्योंकि यह बहुत ठंड है, घास सर्दियों में नहीं उगते।
Compound-Complex sentence	Because grass needs warm weather, it does not grow in winter, but grows in summer.	घास सर्दियों में नहीं उगती, लेकिन गर्मियों में उगती है, क्योंकि इसे गर्म मौसम की जरूरत होती है	क्योंकि घास गर्म मौसम की जरूरत है, यह सर्दियों में नहीं उगते, लेकिन गर्मियों में होती है।

Fig. 8: Sentences translated from Google Translator

The SMT technique is partitioned into four sub-techniques namely word based SMT, phrase-based SMT, syntax based SMT and hierarchy based SMT. these are described below:

Word based SMT: This is the first step in SMT system, in which the given sentence is broken down at the word level and then translation from the SL to a TL is carried out word by word. After this, the re-ordering algorithm is used to arrange the translated words in a meaningful sentence with least error. However, the complexity in the translation arises when compound words like idioms, homonyms comes into picture[16]. Most of the research work in this domain focuses on alignment of translated words. Lambert, De Gispert, Banchs, and Mari~no [18] discussed the guidelines of word alignment evaluation schemes and they also explain how the ratio between ambiguous and unambiguous links in the references impacts the scoring metrics in full-text alignment. Ayan and Dorr [19] observed that higher precision alignments favored phrase-based SMT. Moreover, Lambert, Petitrenaud, Ma, and Way[20] states that PBSMT's BLEU score improves when the alignment is dense and precise in the case of large and small corpora respectively. The author also states that by avoiding long distance crossing links, BLEU score can be improved for small corpora. Och and Ney [21], did well-organized comparison of alignment models with two heuristic models based on Dice coefficient and proposed new statistical alignment model which was a log-linear combination of the models presented in Brown, Della Pietra and Mercer (1993). Avramidis and Koehn[22] provide the solution of translation from morphologically poor language to morphologically rich language by adding per-word linguistic information to the SL.

Phrase-based SMT(PBSMT): This SMT approach was proposed by Koehn, Och, and Marcu[23] with the aim of eliminating the drawbacks of wordbased SMT. In this approach, source and target sentences are separated in phrases(sequence of words), before the translation. Generally, Lexicalized Reordering model is used for the alignment of phrases in the input and output sentences, so that the actual meaning of source sentences will be preserved. Phrase-based SMT performs better than word based SMT due to a fact that memorizing larger unit called phrase, enables it to learn local dependencies such as short distance reordering, insertions and deletions which are internal to the phrase pair[24], but its performance degrades when this technique is subjected to long phrases. Moreover, PBSMT also has some drawbacks such as handling of Nonlocal Dependencies, Weak Reordering Model, Hard Distortion Limit and Spurious Phrasal Segmentation. However researches like cherry[25] minimize the drawback of weak re-ordering, Bisazza and Federico[26] proposed a method for dynamical selection of long range reordering, it helps in increasing the distortion limit up to 18. Tu, Zhou, and Zong[27] proposed the framework that could trace the decoding path on the development set with the optimized translation model and fed the sample training data to the classier. This framework helps in improving the real-time response of MT and reducing the model size while retaining the translation quality and breaks the limitation of specific translation model pruning.

Syntax-based SMT**:** This technique was proposed by Yamada and Knight[28], in which the parse tree of a sentence is subjected to the translation process. The idea of this approach comes under focus after the advent of stochastic parsers in the 1990s. The goal of syntax-based machine translation techniques is to incorporate an explicit representation of syntax into the statistical systems to get the high-quality output without requiring intensive human efforts[29]. Data-oriented parsing based MT and synchronous context-free grammars are the examples of this approach. Like other techniques, this one also has some drawbacks, such as poor scalability and not suitable for languages without a syntactic theory/parser[29]. The parsing problem (cost of decoding) is another drawback of this technique, nevertheless, this can be solved by using expressive translation model which is not compatible with CYK[30]. Furthermore, the quality of syntax-based machine translation greatly depends on semantic of the sentence.

Hierarchical based SMT: Zhang, Huang, Gildea, and Knight [30] proposed this model, with the aim of combining the assets of phrase-based SMT and syntax-based SMT. The main advantage of using hierarchical based SMT is that accuracy of the translation improves due to the higher level of abstraction[16]. Thus it effectively solves the problem of re-ordering of phrases and helps in an incorporation of syntactic information and increasing the efficiency of SMT. To improve the quality and speed of a hierarchical phrase-based SMT system, Yang and Zheng[32] proposed a refined method that combines significant value and compositional properties of surface strings for pruning the phrase table.

SMT has a number of merits over the RBMT[16]which are mentioned below:-

1. SMT does not require any linguistic knowledge whereas RBMT requires linguistic knowledge for the translation.
2. SMT is easy to maintain, can be managed with human supervision and independent from a pair of translating languages.
3. More fluent translations owing to use of a language model.

However, SMT also has some challenges like:-

1. Huge parallel corpus is required, which may not be present every time easily and its creation is little costlier.
2. SMT does not suit the languages with different morphology (eg. Japanese language).
3. It also lacks in grammatical analysis of both SL and TL.
4. Prediction of specific errors in SMT is difficult.

With these pros and cons, SMT suits well for the machine translation of English to European languages. Google Translate and Microsoft Translator are the platforms which use SMT for translation.

Example based machine translation(EBMT): EBMT was proposed by Makoto Nargo in 1981 with an aim of translation by analogy. EBMT is defined as a system having corpora of sentences in the SL and its translation in the TL with a point to point mapping. It can be viewed as an implementation of a case-based reasoning approach to machine learning, which means solving newer problems based on the solution of similar past problems. This system is preferred for the translation between two different languages like English and Japanese, and this is the major asset of EBMT over SMT. Its translation approach is very similar to the method, through which a human translates a sentence, that is, first, the sentence is disintegrated into phrases, then these phrases are translated separately and finally long translated sentence is formed by the proper alignment of translated phrases. Thus, without going into the deep linguistic analysis(as required in RBMT) translation of sentences are possible in EBMT.

Example of English to Hindi translation in EBMT system:-

Input sentence in the English language = "Shubham wants to become a bureaucrat"
Output sentence in the Hindi language = “Shubham ek naukarashaa banana chahata hain"

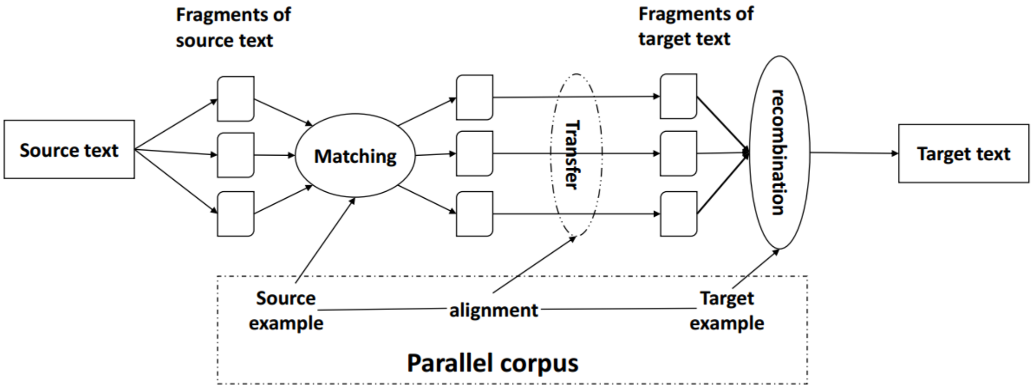


Fig. 9: Example -Based Machine Translation

The translation methodology of EBMT(as shown in fig. 9) is explained by considering an above example. First, the matching stage comes in which a source sentence(sentence in English) is partitioned into several fragments like “want to", “bureaucrat", “shubham", etc, followed by examples are searched from parallel corpora based on some similarity measures like word similarity or syntactic and semantic similarity. Somers[33] explains various matching techniques for translation like partial matching, structured based matching, character based matching etc. Then, in transfer stage, TL fragments corresponding to relevant fragments are taken out with its translation in TL like system selects “Shita wants to become a bureaucrat" with its translation “shita ek naukarashaah banana chaahati hain" from corpora. After this, recombination stage come in which, phrases are organized in a meaningful manner as per the source sentence and , here ‘shubham’ is replaced with ‘shita’ and ‘chaahata’ with ‘chaahati’ and finally, after recombination translated sentence is generated.

Advantages of EBMT are:-

1. This technique avoids the use of manually driven rules.
2. Minimum prior knowledge is required in EBMT which makes it an attractive system.
3. EBMT system is adaptable to many language pairs.

Disadvantages of EBMT are:-

1. The efficiency of EBMT system may hampered in case of noisy corpora, because efficient technique is not available to clean the noisy corpora.
2. Efficient computational model requires for large database.

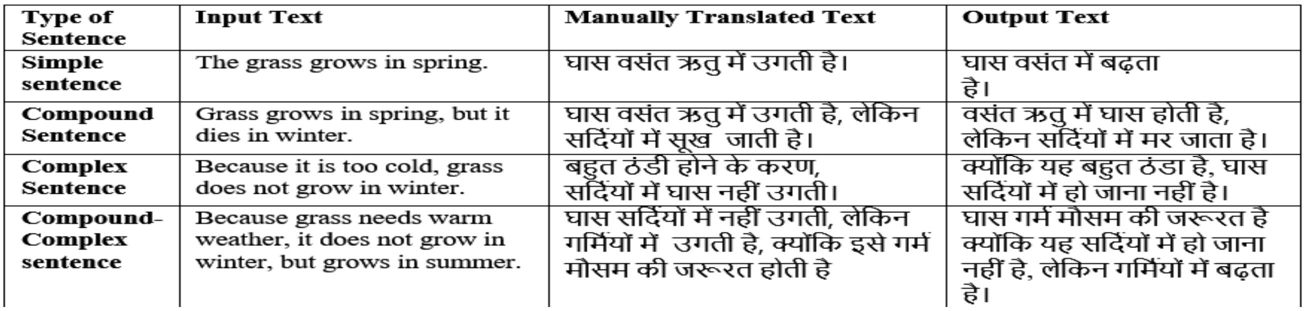
However, having such pros and cons, some systems like ALEPH, wEBMT, and PanEBMT uses example-based MT approach and it is preferred for English to Turkish, English to Japanese translation.

TABLE III Comparison of SMT and EBMT

S.No.	SMT	EBMT
1	SMT is based on statistical models whose parameters are derived from the analysis of bilingual text corpora.	EBMT is based on the use of bilingual corpus with parallel texts, in which translation by the analogy is the main ideology.
2	When SMT encountered previously seen data, it treats as unseen/ new data. Means SMT does not learn from the data that come across it.	EBMT simply search its bilingual corpora for that string and translate it directly. Thus EBMT learns from the data that come across it.
3	SMT alignments are not available for reuse in the system, as alignments are disappeared in the probability models.	EBMT alignments are available for reuse in the system

C. Hybrid-based Machine Translation

The hybrid-based Machine translation (HMT) was introduced with an aim of eliminating the demerits of former above-mentioned approaches by developing a single machine translation system that utilizes merits of various translation approaches. The motivation for developing hybrid machine translation systems stems from the failure of any single technique to achieve a satisfactory level of accuracy. The basic ideology behind the HMT is that first the source sentence is modified by using any translation technique(say RBMT), then the modified sentence is subjected to some other translation technique(say SMT) so as to get more refined output. In this way, the use of RBMT removes the drawbacks of SMT like difficulty in matching of source sentence with target sentence present in corpora. Whereas the use of SMT removes the drawback of RBMT like Failure to adapt to new domains. In fig. 10, we provide translations of Microsoft Bing translator (http://www.bing.com/translator), which uses HMT for translation, in the figure we mentioned the translation of different type of English sentences into a corresponding Hindi sentences. Moreover, figure 10 also compares machine translated text from manually translated text.



Type of Sentence	Input Text	Manually Translated Text	Output Text
Simple sentence	The grass grows in spring.	घास वसंत ऋतु में उगती है।	घास वसंत में बढ़ता है।
Compound Sentence	Grass grows in spring, but it dies in winter.	घास वसंत ऋतु में उगती है, लेकिन सर्दियों में सूख जाती है।	वसंत ऋतु में घास होती है, लेकिन सर्दियों में मर जाता है।
Complex Sentence	Because it is too cold, grass does not grow in winter.	बहुत ठंडी होने के करण, सर्दियों में घास नहीं उगती।	क्योंकि यह बहुत ठंडा है, घास सर्दियों में हो जाना नहीं है।
Compound-Complex sentence	Because grass needs warm weather, it does not grow in winter, but grows in summer.	घास सर्दियों में नहीं उगती, लेकिन गर्मियों में उगती है, क्योंकि इसे गर्म मौसम की जरूरत होती है	घास गर्म मौसम की जरूरत है क्योंकि यह सर्दियों में हो जाना नहीं है, लेकिन गर्मियों में बढ़ता है।

Fig. 10: Sentences translated from Microsoft Bing Translator

Numerous models are proposed for the hybridization of different techniques, which can be classified into Multi-Engine model, statistical rule generation model, Statistics Guided by Rules model and Multi-pass model. Each of them are explained below:-

Multi-Engine Model: Under this kind of hybridization, multiple systems based on different MT techniques are used in parallel, but the final output is produced by combining the output of all the systems. The output quality of this type of translation heavily depends on the type of the technique used, and the manner in which different intermediate results are combined to produce a final result. Various systems has been proposed like system that combines SMT and RBMT with an Open-Source Decoder in multi-engine setup[9] Moreover, hybrid architecture of multiple RBMT using Moses was proposed, that works better that SMT with increased lexical coverage[34]. However, this architecture has some inherent drawback such as it requires more computational efforts because decoder generates many instances of an identical results due presence of multiple instances of the same phrase pair in combined phrase table.

Statistical rule generation model: This model is basically an RBMT system which is constructed by using corpora. Under this approach, statistical data is used to generate lexical and syntactic rules, then these rules are used to translate source sentence in rule-based translator[35]. Moreover, sometimes rules/dictionary are enriched by either extracting phrases and examples from parallel corpora or extracting new entities from BabelNet and Wiktionary. The main advantage of such hybridization is that, it requires less time, efforts and cost for translation. However, this approach has some demerits like the accuracy of the translation depends heavily on the similarity of the input text and the text of the training corpus. Thus, it is suitable for domain-specific applications.

*Statistics Guided by Rules model***:** The basic ideology is that, rules are used to reorder the given sentence to make it more suitable for DDMT. In this model the rules can be applied either at the pre-processing stage or post-processing stage or at the core model of the system[35]. Rules at the pre-processing are generally intended to re-arrange the source sentence in such a way that, source sentence matches better with the target sentence [36]. Whereas rules at the post-processing are intended to generate morphology by an introduction of dictionaries and machine learning. The normalization problem of a noisy channel is solved by the combination of pre and post processing. Whereas rules are introduced at the core of the system to improve statistical word alignment, to integrate morphology and syntax knowledge dynamically in the system[8]. This kind of hybridization provides more flexibility and control over translation.

Multi-pass model: The main idea behind this approach is to eliminate the demerits of the deployed technique by using the technique other than deployed one in further processing. The most common multi-pass machine translation system is statistical smoothing or automatic post editing system, in which the input sentence is pre-processed with a rule-based machine translation system to generate intermediate output. Then, the obtained intermediate output is processed with statistical machine translation system, so as to get the final output[37]. This resolve the problems of lack of analysis in RBMT, failure to adopt new domain, etc. As a result, the overall quality of translation improves significantly(in term of BLUE score). The biggest advantage of this approach is the ability to deal with ambiguity of translation, which is the greatest challenge of RBMT. When a word/phrases have more than one meaning, statistics helps to identify the most suitable option. Contrary to that, HMT is unable to curtain presence of implicit limitations that are inherited from techniques involved like the high cost of RBMT is still present in the hybrid model. Moreover, it also introduces additional complexities of managing side-by-side systems making their true commercial value questionable. However, HMT is successfully implemented in PROMT, SYSTRAN and Asia Online translating platform.

V. Conclusion

In this paper, an overview of proposed techniques is presented in a classified manner, along with their respective research work. With the analysis of these techniques, we found that translation quality has been improved a lot, especially after the advent of hybrid machine translation technique. Moreover, the research in the field of removing ambiguities up to a certain level, improving fluency, increasing reusability and reducing dependency on both resources(linguistic knowledge) and pair of language(SL and TL), contributes a lot in improving the quality of translation. But despite of these efforts quality of fully automated translation is still not superior to that of fully manual translation because of requirement of the big dictionaries or corpora, complexity of system that hampers its commercial viability, domain specific nature of system, lexicon and linguistic irregularities, etc.

The importance of our finding is that quality of translation can be improved by proper analysis of SL, that involves study of grammar, semantic, morphology and syntactic. Moreover quality can also be improved by completely eliminating the ambiguity of the involved languages and deep understanding of TL. Thus, for the future work, we can conclude that, translation through meta-phrase is possible but through para-phrase is still difficult and requires more research to explore it better. Moreover, problem of ambiguity in the words/phrases also requires more efforts.

References

1. Ch´eragui, Mohamed Amine (2012), “Theoretical overview of machine translation." Proceedings ICWIT, 160.
2. Costa-Jussa, Marta R, Mireia Farr´us, Jos´e B Marino, and Jos´e AR Fonollosa (2012), “Study and comparison of rule-based and statistical catalan-spanish machine translation systems." Computing and informatics, 31, 245-270.
3. Brown, Peter F, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Rober Mercer, and Paul S Roossin (1990),”A statistical approach to machine translation." Computational linguistics, 16, 79-85.
4. Slocum, Jonathan (1985), “A survey of machine translation: its history, current status, and future prospects." Computational linguistics, 11, 1-17.
5. Hutchins, W John (1988), “Recent developments in machine translation a review of the last five years." In Presented at conference on ‘New directions in machine translation, volume 18, 19, Citeseer.
6. Dorr, Bonnie J, Pamela W Jordan, and John W Benoit (1999), “A survey of current paradigms in machine translation. Advances in computers, 49, 1-68.
7. Tripathi, Sneha and Juran Krishna Sarkhel (2010), “Approaches to machine translation." Annals of library and information studies, 57, 388-393
8. Antony, PJ (2013), “Machine translation approaches and survey for indian languages." Computational Linguistics and Chinese Language Processing, 18, 47-78.
9. Okpor, MD (2014), “Machine translation approaches: issues and challenges." International Journal of Computer Science Issues (IJCSI), 11, 159.
10. Kituku, Benson, Lawrence Muchemi, and Wanjiku Nganga (2016), “A review on machine translation approaches." Indonesian Journal of Electrical Engineering and Computer Science, 1, 182-190.
11. Charoenpornsawat, Paisarn, Virach Sornlertlamvanich, and Thatsanee Charoenporn (2002), “Improving translation quality of rule-based machine translation." In Proceedings of the 2002 COLING workshop on Machine translation in AsiaVol, 16, 1-6, Association for Computational Linguistics.
12. Kaji, Hiroyuki (1988), “An efficient execution method for rule-based machine translation." In Proceedings of the 12th conference on Computational linguistics Volume 2, 824-829, Association for Computational Linguistics.
13. Ch´eragui, Mohamed Amine (2012), “Theoretical overview of machine translation." Proceedings ICWIT, 160.
14. Hatem, Arwa, Nazlia Omar, and Khalid Shaker (2011), “Morphological analysis for rule based machine translation." In 2011 International Conference on Semantic Technology and Information Retrieval, 260-263, IEEE.
15. Dave, Shachi, Jignashu Parikh, and Pushpak Bhattacharyya (2001), “Interlinguabased english{hindi machine translation and language divergence." Machine Translation, 16, 251-304.
16. Chen, Boxing and Marcello Federico (2006), “Improving phrase-based statistical translation through combination of word alignments." In Advances in Natural Language Processing, 356-367, Springer.
17. Lopez, Adam (2008), “Statistical machine translation." ACM Computing Survey (CSUR), 40, 8.
18. Lambert, Patrik, Adri`a De Gispert, Rafael Banchs, and Jos´e B Mari~no (2005), “Guidelines for word alignment evaluation and manual alignment." Language Resources and Evaluation, 39, 267-285
19. Ayan, Necip Fazil and Bonnie J Dorr (2006), “Going beyond aer: An extensive analysis of word alignments and their impact on mt." In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 9-16 Association for Computational Linguistics.
20. Lambert, Patrik, Simon Petitrenaud, Yanjun Ma, and Andy Way (2012), “What types of word alignment improve statistical machine translation?" Machine translation, 26, 289-323.
21. Och, Franz Josef and Hermann Ney (2003), “A systematic comparison of various statistical alignment models." Computational linguistics, 29, 19-51
22. Avramidis, Eleftherios and Philipp Koehn (2008), “Enriching morphologically poor languages for statistical machine translation." In ACL, 763-770.
23. Koehn, Philipp, Franz Josef Och, and Daniel Marcu (2003), “Statistical phrasebased translation." In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, 48-54, Association for Computational Linguistics.
24. Durrani, Nadir, Helmut Schmid, Alexander Fraser, Philipp Koehn, and Hinrich Sch¨utze (2015), “The operation sequence model combining n-gram-based and phrase-based statistical machine translation." Computational Linguistics.
25. Lagarda, A-L, Vicente Alabau, Francisco Casacuberta, Roberto Silva, and Enrique Diaz-de Liano (2009), “Statistical post editing of a rule-based machine translation system." In Proceedings of Human Language Technologies: The 2009 Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, 217-220, Association for Computational Linguistics
26. Bisazza, Arianna and Marcello Federico (2013), “Efficient solutions for word reordering in german-english phrase-based statistical machine translation." In Proceedings of the Eighth Workshop on Statistical Machine Translation, 440-451.
27. Tu, Mei, Yu Zhou, and Chengqing Zong (2015), “Exploring diverse features for statistical machine translation model pruning." IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23, 1847-1857.
28. Liu, Lemao and Liang Huang (2014), “Search-aware tuning for machine translation." In EMNLP, 1942-1952.
29. Ahmed, Amr and Greg Hanneman (2005), “Syntax-based statistical machine translation: A review." In *Association for Computational Linguistics*.

30. Galley, Michel, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer (2006), “Scalable inference and training of context-rich syntactic translation models." In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 961-968, Association for Computational Linguistics.
31. Zhang, Hao, Liang Huang, Daniel Gildea, and Kevin Knight (2006), “Synchronous binarization for machine translation." In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 256{263, Association for Computational Linguistics.
32. Yang, Mei and Jing Zheng (2009), “Toward smaller, faster, and better hierarchical phrase-based smt." In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, 237-240, Association for Computational Linguistics.\
33. Somers, Harold (1999), “Review article: Example-based machine translation." Machine Translation, 14, 113-157.
34. Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. (2007), “Moses: Open source toolkit for statistical machine translation." In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, 177-180, Association for Computational Linguistics.
35. Costa-Jussa, Marta R and Jos´e AR Fonollosa (2015), “Latest trends in hybrid machine translation and its applications.Computer Speech & Language, 32, 3-10.
36. Collins, Michael, Philipp Koehn, and Ivona Kuˇcerov´a (2005), “Clause restructuring for statistical machine translation." In Proceedings of the 43rd annual meeting on association for computational linguistics, 531-540, Association for Computational Linguistics.
37. Lagarda, A-L, Vicente Alabau, Francisco Casacuberta, Roberto Silva, and Enrique Diaz-de Liano (2009), “Statistical post editing of a rule-based machine translation system." In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, 217-220, Association for Computational Linguistics.

AUTHOR(S) PROFILE

Shantanoo Dubey, Currently, pursuing Bachelor of Engineering in Computer Science and Engineering from Jabalpur Engineering College during 2013-17.